International Journal of Medical Informatics 102 (2017) 138–149

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.ijmijournal.com

A method for cohort selection of cardiovascular disease records from an electronic health record system

Maria Tereza Fernandes Abrahão[a,*], Moacyr Roberto Cuce Nobre[b], Marco Antonio Gutierrez[c]

[a] Program in Cardiology, Heart Institute (InCor) Clinical Hospital, Faculty of Medicine, University of Sao Paulo, Sao Paulo, Brazil
[b] Clinical Epidemiology Team, Heart Institute (InCor) Clinical Hospital, Faculty of Medicine, University of Sao Paulo, Sao Paulo, Brazil
[c] Biomedical Informatics Laboratory, Heart Institute (InCor) Clinical Hospital, Faculty of Medicine, University of Sao Paulo, Sao Paulo, Brazil

ARTICLE INFO

Article history:
Received 11 March 2016
Received in revised form 10 February 2017
Accepted 24 March 2017

Keywords:
Secondary use of data
Medical informatics
Electronic health records
Cohort studies
Retrospective studies and data mining

ABSTRACT

Introduction: An electronic healthcare record (EHR) system, when used by healthcare providers, improves the quality of care for patients and helps to lower costs. Information collected from manual or electronic health records can also be used for purposes not directly related to patient care delivery, in which case it is termed secondary use. EHR systems facilitate the collection of this secondary use data, which can be used for research purposes like observational studies, taking advantage of improvement in the structuring and retrieval of patient information. However, some of the following problems are common when conducting a research using this kind of data: (i) Over time, systems and data storage methods become obsolete; (ii) Data concerns arise since the data is being used in a context removed from its original intention; (iii) There are privacy concerns when sharing data about individual subjects; (iv) The partial availability of standard medical vocabularies and natural language processing tools for non-English language limits information extraction from structured and unstructured data in the EHR systems. A systematic approach is therefore needed to overcome these, where local data processing is performed prior to data sharing.
Method: The proposed study describes a local processing method to extract cohorts of patients for observational studies in four steps: (1) data reorganization from an existing local logical schema into a common external schema over which information can be extracted; (2) cleaning of data, generation of the database profile and retrieval of indicators; (3) computation of derived variables from original variables; (4) application of study design parameters to transform longitudinal data into anonymized data sets ready for statistical analysis and sharing. Mapping from the local logical schema into a common external schema must be performed differently for each EHR and is not subject of this work, but step 2, 3 and 4 are common to all EHRs. The external schema accepts parameters that facilitate the extraction of different cohorts for different studies without having to change the extraction algorithms, and ensures that, given an immutable data set, can be done by the idempotent process. Statistical analysis is part of the process to generate the results necessary for inclusion in reports. The generation of indicators to describe the database allows description of its characteristics, highlighting study results. The set extraction/statistical processing is available in a version controlled repository and can be used at any time to reproduce results, allowing the verification of alterations and error corrections. This methodology promotes the development of reproducible studies and allows potential research problems to be tracked upon extraction algorithms and statistical methods
Results: This method was applied to an admissions database, SI^3, from the InCor-HCFMUSP, a tertiary referral hospital for cardiovascular disease in the city of São Paulo, as a source of secondary data with 1116848 patients records from 1999 up to 2013. The cleaning process resulted in 313894 patients records and 27698 patients in the cohort selection, with the following criteria: study period: 2003–2013, gender: Male, Female, age: ≥ 18 years old, at least 2 outpatient encounters, diagnosis of cardiovascular disease (ICD-10 codes: I20-I25, I64-I70 and G45). An R script provided descriptive statistics of the extracted cohort.

* Corresponding author.
E-mail addresses: tereza.abrahao@usp.br, terezaabrahao2012@gmail.com (M.T.F. Abrahão).

http://dx.doi.org/10.1016/j.ijmedinf.2017.03.015

Conclusion: This method guarantees a reproducible cohort extraction for use of secondary data in observational studies with enough parameterization to support different study designs and can be used on diverse data sources. Moreover it allows observational electronic health record cohort research to be performed in a non-English language with limited international recognized medical vocabulary.

1. Introduction

Healthcare databases of electronic healthcare record (EHR) systems are a source of data, which favour observational studies and contain the necessary structures for clinical research. Prospective and retrospective studies have been developed for scientific use of this information, also known as secondary use because data is used in a context different from its original intention.

There are a number of methodological challenges encountered when working with this data [1]. Among the advantages of health care databases for research are their low cost and the possibility of using larger sample sizes, providing the ability to detect small differences or rare events and avoiding direct contact with patients. In addition they provide greater methodological diversity, allowing changes in the study to be implemented differently without the need for rigorous protocols required in research projects such as randomized trials. The disadvantages are related to the lack of standardization in data collection, which affects the quality of the recorded data, changes in data collection procedures over time, and the lack of specific information that may be important for analysis, including outcome variables, explanations, mediators, or presence of confusion [2].

This article describes a cohort extraction method from a healthcare database for use in observational studies. In this case the database is from the InCor-HCFMUSP, a tertiary referral hospital for cardiovascular disease in the city of São Paulo, known as the SI^3 EHR system database [3,4]. The method is based on reproducibility criteria and is validated by the application of a study on cardiovascular disease (CVD) (diagnosis ICD-10 [5] codes: I20 to I25, I64 to I70, G45) and statin medication.

2. Method

The proposed method consists in applying algorithms based upon relational algebra, represented by instructions in standard language ANSI SQL99 (ISO/IEC 9075: 1999), which select and map information to an external schema, and systematically clean, process and extract a data set for use in observational studies.

The SI^3 primary healthcare database used as source of secondary use data, conforms to the essential properties required for the methodology to be applied: (i) the health care system information is stored in a relational database [6,7]; (ii) there is a unique identifier for the patient record.

The cohort extraction method is represented as an external schema that, through successive applications of relations (views), allows systematic extraction of cohorts without manual intervention. It is idempotent, meaning that it will generate the same results if repeated, without changing the state of the database or the original data. A set of variables is proposed for mapping the logical model of the database to the external schema on which the method is applied. Mapped data is applied to other views for cleaning and processing of data. The method uses parameters that define the cohort (start date, end of study date, diagnosis, intervention, outcome, etc.) and performs statistical analysis on the data resulting from the extraction of the cohort by applying a set of functions and algorithms using statistical software. The database is profiled to produce indicators of data quality and database characteristics, for example percentage of duplicates, number of null values, and statistics about predominant diseases in hospital care (distribution diagnostics). A determined cohort extracted by the method, together with the values of parameters used, SQL code and statistical analysis code is stored in a version-controlled repository with the version number uniquely defining the extracted cohort.

2.1. Application of the method

The external schema includes data cleaning, processing and consolidation. Study design criteria were applied for extraction of cohort and statistical analysis. Fig. 1 shows the method for cohort extraction and analysis applied to the source data, which is described in the followings sections.

3. External schema

The external layout consists of a set of views on the model, as shown in Fig. 2, where the patient view is seen in relation to various other views. The relationship of the patient view to death is always one-to-one, whereas all other views have a one-to-many relationship. The set of views is created once for each source database. The method depends only on this set of views and mapped information.

Fig. 3 shows an example mapping to the SI^3 database, which is specific for each database and above the level at which the method is applied. The mapping must be performed before the application of the method.

Various tables represent patient data in the SI^3 database. Personal patient data, demographic and identifying data, and patient records are mapped to the external patient view. Diagnostic data and admission dates were mapped to the external diagnosis view.

4. Cleaning method

Fig. 4 shows in detail the views that comprise the method of cohort cleaning and screening, using as an example the relationship between Patient and Diagnosis views. Patients who did not have a diagnosis were not included in the study. The various stages involved in cohort extraction, enrollment filters, inconsistency detectors, intermediate database stage, cohort selectors and transformers are also shown.

The enrollment filter stage allows selection of records according to a specific and independent feature of the study. A time window (starting and ending date) is defined in which records should be retrieved, ensuring that a subsequent repetition of the process at a different date retrieves the same record set. Characterization of the database is performed by a statistical analysis of the records that pass the enrollment filter.

In the data cleaning stage inconsistency detectors are applied, disposing of records that have variables with some inconsistency, e.g. void identification fields, invalid dates, and duplicate records. The outputs of the detectors are used to compute quality indicators. The results from the clean records view form the intermediary database stage.